A FERGUSON-ANTONIAK APPROACH TO THE EMPIRICAL BAYES ESTIMATION OF A BINOMIAL PARAMETER

by

Ronald Christensen

Technical Report No. 271

January, 1977

A FERGUSON-ANTONIAK APPROACH TO THE EMPIRICAL BAYES ESTIMATION OF A BINOMIAL PARAMETER

by

Ronald Christensen

1. Introduction

The random variables X_i, $i = 1, \ldots, k$, are binomially distributed with parameters n_i and θ_i. The values n_i are known and $\theta_1, \ldots, \theta_k$ are chance observations from a probability distribution G which concentrates its mass on $[0, 1]$. The measure G and the values of $\theta_1, \ldots, \theta_k$ are unknown. The empirical Bayes problem is to make inferences about G or a particular θ_i using the information from $(X_1, \ldots, X_k)$. Most existing procedures, see [5], for estimating θ_k, say, first use $X_1, \ldots, X_{k-1}$ and $n_1, \ldots, n_{k-1}$ to estimate G, then they apply Bayes' theorem using (x_k, n_k) to obtain the posterior distribution of θ_k:

$$dH(\theta_k | x_k, n_k) \propto \theta_k^{x_k}(1 - \theta_k)^{n_k - x_k} dG(\theta_k).$$

The approach taken here is in the mainstream of Bayesian statistics, rather than being empirical Bayesian, in that G is a random probability measure, the distribution of which is modified by $(x_1, n_1, \ldots, x_k, n_k)$ in making inferences about G or θ_k, say, where in the latter case (x_k, n_k) plays a special role.

In Section 2 a review of the Dirichlet process will be given,

Section 3 contains a review of mixtures of Dirichlet processes. Section 4 will give the results for the binomial application discussed here.

2. Dirichlet processes.

In his 1973 paper [4], Ferguson considers a probability space $(\Theta, \mathfrak{F})$ and defines a probability measure P on the space $([0, 1]^{\mathfrak{F}}, \mathbb{B}\mathfrak{F})$, where $[0, 1]^{\mathfrak{F}}$ is the collection of set functions from $\mathfrak{F}$ to $[0, 1]$ and $\mathbb{B}\mathfrak{F}$ is the σ-field generated by the cylinder sets. This is done by defining a stochastic process P indexed by sets in $\mathfrak{F}$. If the finite-dimensional distributions of such a process can be defined consistently we can, in the usual way (see Breiman [3], Chapter 12, Section 3), show its existence. The standard proof is not affected by the unusual nature of the index set.

If for any n and measurable partition of Θ, $\{B_1, \ldots, B_n\}$, $(P(B_1), \ldots, P(B_n))$ has a Dirichlet distribution with parameters $\alpha(B_1), \ldots, \alpha(B_n)$, where α is a finite non-null measure on $(\Theta, \mathfrak{F})$--written $(P(B_1), \ldots, P(B_n)) \sim \mathcal{D}(\alpha(B_1), \ldots, \alpha(B_n))$--then it can be shown ([4], Lemma 1) that the finite-dimensional distributions can be defined consistently. We call the process P (which is defined more completely in Ferguson [4]) a Dirichlet process with parameter α and write $P \sim \mathcal{D}(\alpha)$.

To see that this process defines a random probability distribution on the space $(\Theta, \mathfrak{F})$ it is necessary to show that P gives no weight to set functions in $[0, 1]^{\mathfrak{F}}$ that are not probability measures. If $\emptyset$ is the empty set, then $P(\Theta) = 1$ a.s. and $P(\emptyset) = 0$ a.s. as can be seen from properties of the Dirichlet distribution. That P is almost surely countably additive can be shown by application

of the Borel-Cantelli lemma (see [4], Proposition 2).

Alternatively, Ferguson defines a process $\tilde{P}$, indexed by $\mathfrak{F}$, on a probability space $(\Omega, \mathfrak{B})$ such that for any $A \in \mathfrak{F}$ and any $\omega \in \Omega$

$$\tilde{P}(A)(\omega) = \sum_{j=1}^{\infty} p_j(\omega)\delta_{V_j(\omega)}(A),$$

where p_j and V_j are as follows. Let α be a finite non-null measure on $(\Theta, \mathfrak{F})$. Each V_j is a measurable function from $(\Omega, \mathfrak{B})$ to $(\Theta, \mathfrak{F})$; the V_j's are i.i.d. with probability distribution $\frac{\alpha(\cdot)}{\alpha(\Theta)}$ and they are independent of the p_j's. The p_j's are random variables which depend on $\alpha(\Theta)$ and have the property that $\sum_{j=1}^{\infty} p_j = 1$ a.s.; $\delta_x(\cdot)$ is a probability measure on $(\Theta, \mathfrak{F})$ giving mass one to the point x.

$\tilde{P}$ is a measurable mapping of $(\Omega, \mathfrak{B})$ to $([0, 1]^{\mathfrak{F}}, \mathcal{B}\mathfrak{F})$ and hence induces a probability measure $\tilde{\mathcal{P}}$ on $([0, 1]^{\mathfrak{F}}, \mathcal{B}\mathfrak{F})$ in the usual way. Ferguson shows that the process $\tilde{P}$ has the same distribution as the process P so by the uniqueness part of the Kolmogorov extension theorem (see Ash[2]), $\mathcal{P} = \tilde{\mathcal{P}}$. Since it is clear that $\tilde{P}(\cdot)$ is almost surely a discrete probability measure, $\tilde{\mathcal{P}}$ must give all its mass to discrete probability measures. Since $\mathcal{P} = \tilde{\mathcal{P}}$, $P(\cdot)$ must be a discrete probability measure on $(\Theta, \mathfrak{F})$ a.s.

$(\theta_1, \ldots, \theta_k)$ is defined to be a sample of size k from P if for any positive integers m and k and any $\mathfrak{F}$-measurable sets $A_1, \ldots, A_m, C_1, \ldots, C_k$.

$$\wp\{\theta_1 \epsilon C_1, \ldots, \theta_k \epsilon C_k | P(A_1), \ldots, P(A_m), P(C_1), \ldots, P(C_k)\}$$

$$= \prod_{j=1}^{k} P(C_j) \text{ a.s.}$$

Since the distribution of P is known this defines a probability $\wp$ on $(\Theta^k \times [0, 1]^{\mathfrak{F}}, \mathfrak{F}^k \times \mathbb{B}^{\mathfrak{F}})$ through the Kolmogorov extension. Theorem 1 in [4] states that if $P \sim \mathfrak{D}(\alpha)$ and $(\theta_1, \ldots, \theta_k)$ is a sample of size k from P then

$$P|(\theta_1, \ldots, \theta_k) \sim \mathfrak{D}(\alpha + \sum_{i=1}^{k} \delta_{\theta_i}).$$

Another useful property is that if $P \sim \mathfrak{D}(\alpha)$ then

$$\wp(\theta \epsilon A) = E[\wp(\theta \epsilon A \,|P(A))] = EP(A) = \frac{\alpha(A)}{\alpha(\Theta)}.$$

The last equality follows because $P(A)$ has a beta distribution.

We demonstrate how these facts apply to statistical problems with two examples that were considered by Ferguson. Set $(\Theta, \mathfrak{F}) = (\mathbb{R}, \mathbb{B})$ and assume that P has a Dirichlet process prior distribution on $(\mathbb{R}, \mathbb{B})$ with parameter α. A random sample $(\theta_1, \ldots, \theta_k)$ is taken from P so the posterior distribution of P is $\mathfrak{D}(\alpha + \sum_{i=1}^{k} \delta_{\theta_i})$. Suppose the decision problem is to estimate the distribution function G with loss function $L(P, \hat{G}) = \int (G(t) - \hat{G}(t))^2 \, dW(t)$, where $G(t) = P((-\infty, t])$, $\hat{G}(t)$ is an estimate of $G(t)$, and W is a finite weighting measure on $(\mathbb{R}, \mathbb{B})$. A Bayes estimate of G is, for any W,

$$\hat{G}(t) = \frac{\alpha(\mathbb{R})}{\alpha(\mathbb{R}) + k} \frac{\alpha((-\infty, t])}{\alpha(\mathbb{R})} + \frac{k}{\alpha(\mathbb{R}) + k} F_k(t|\theta_1, \ldots, \theta_k) \quad ,$$

where $F_k(t|\theta_1, \ldots, \theta_k)$ is the empirical distribution function. The Bayes estimate is a convex combination of the optimal prior estimate $\frac{\alpha((-\infty, t])}{\alpha(\mathbb{R})}$ and the empirical c.d.f. The weight associated with the prior estimate is proportional to $\alpha(\mathbb{R})$ which therefore serves as the effective number of "prior observations." If the decision problem is to estimate the mean μ of P with $\hat{\mu}$ and loss function $L(P, \hat{\mu}) = (\mu - \hat{\mu})^2$ then the Bayes estimate is $\hat{\mu} = \frac{\alpha(\mathbb{R})}{\alpha(\mathbb{R}) + k} \mu_0 + \frac{k}{\alpha(\mathbb{R}) + k} \bar{\theta}$ where μ_0 is the mean of $\frac{\alpha(\cdot)}{\alpha(\mathbb{R})}$, assumed to exist, and $\bar{\theta}$ is the sample mean. This estimate is simply the mean of the Bayes estimate $\hat{G}$ from the previous example.

3. Mixtures of Dirichlet processes.

In the problem posed in Section 1, let the prior distribution of the probability measure be P, a Dirichlet process with parameter α, then, as shown by Antoniak [1], the posterior distribution is a mixture of Dirichlet processes.

To consider a mixture of Dirichlet processes it is necessary to extend the notion of the parameter of the process. For two measurable spaces $(\Theta, \mathcal{G})$ and $(U, \mathcal{B})$ define a transition measure on $U \times \mathcal{G}$ to be a mapping $\alpha: U \times \mathcal{G} \to [0, \infty)$ such that

1) for all $u \in U$, $\alpha(u, \cdot)$ is a finite non-null measure on $(\Theta, \mathcal{G})$,

2) for all $A \in \mathcal{G}$, $\alpha(\cdot, A)$ is measurable w.r.t. $\mathcal{B}$.

If H is a probability on $(U, \mathcal{B})$ and α is a transition measure we can define a mixture of Dirichlet processes with these parameters by specifying for a measurable partition of Θ, $\{B_1, \ldots, B_n\}$ that

$$P\{P(B_1) \leq y_1, \ldots, P(B_n) \leq y_n\}$$

$$= \int_U D(y_1, \ldots, y_n | \alpha(u, B_1), \ldots, \alpha(u, B_n)) dH(u)$$

where $D(y_1, \ldots, y_n | \alpha_1, \ldots, \alpha_n)$ is the Dirichlet distribution function with parameters $\alpha_1, \ldots, \alpha_n$.

In accordance with previous notation write

$(P(B_1), \ldots, P(B_n)) \sim \int_U \mathcal{D}(\alpha(u, B_1), \ldots, \alpha(u, B_n))dH(u)$ and $P \sim \int_U \mathcal{D}(\alpha(u, \cdot))dH(u)$. Clearly, conditional on u, P is a Dirichlet process, written $P|u \sim \mathcal{D}(\alpha_u)$. The definition of a random sample from P must be modified so that

$$\mathcal{P}\{\theta_1 \in C_1, \ldots, \theta_k \in C_k | u, P(A_1), \ldots, P(A_m), P(C_1), \ldots, P(C_k)\}$$

$$= \prod_{j=1}^{k} P(C_j).$$

For measurable spaces $(\Theta, \mathcal{C})$, $(X, \mathcal{C})$, $(U, \mathcal{B})$, with some mild regularity conditions, Antoniak [1] proves the following theorem. Let α be a transition measure on $U \times \mathcal{C}$, and F a transition probability measure on $\Theta \times \mathcal{C}$. Let H be a probability on $(U, \mathcal{B})$, $P \sim \int_U \mathcal{D}(\alpha_u)dH(u)$ and θ is a sample from P. If X is a random variable whose distribution, conditional on P, u, and θ, is $F(\theta, \cdot)$ then

$$P|X \sim \int_{\Theta \times U} \mathcal{D}(\alpha_u + \delta_\theta)dH_X(\theta, u);$$

where $H_X(\theta, u)$ is the distribution of (θ, u) given X. That is, if $u \sim H$, $P|u \sim \mathcal{D}(\alpha_u)$, $P \sim \int_U \mathcal{D}(\alpha_u)dH(u)$, $\theta|(P, u) \sim P$ and $X|(P, \theta, u) \sim F(\theta, \cdot)$, then

$$P|X \sim \int_{\Theta \times U} \mathcal{D}(\alpha_u + \delta_\theta)dH_X(\theta, u).$$

As a corollary to the above theorem, if $P \sim \mathcal{D}(\alpha)$, $\theta_1, \ldots, \theta_k$ is a sample from P and the X_i's are random

variables such that

$$(X_1, \ldots, X_n)|(P, \theta_1, \ldots, \theta_k) \approx \prod_{i=1}^{k} F(\theta_i, X_i)$$

then since θ_1 is a sample from P and for $i=2, \ldots, k$, θ_i is a sample from $P|(\theta_1, \ldots, \theta_{i-1})$

$$P|(X_1, \ldots, X_n) \sim$$
$$\int_{\Theta^k} \mathcal{D}(\alpha + \delta_{\theta_1} + \ldots + \delta_{\theta_k}) dH_{X_1, \ldots, X_k}(\theta_1, \ldots, \theta_k)$$

where $H_{X_1, \ldots, X_k}(\theta_1, \ldots, \theta_k)$ is the distribution of $(\theta_1, \ldots, \theta_k)$ given $(X_1, \ldots, X_k)$.

The corollary resolves several statistical problems. If it is desired to estimate θ_k with squared error loss then a Bayes estimate is $E(\theta_k|X_1, \ldots, X_k)$. If the problem is to estimate the distribution of $\theta \epsilon \Theta$ with the loss function $L(P, \hat{G}) = \int_{-\infty}^{\infty} (G - \hat{G})^2 dW$ and notation as in Section 2 then a Bayes estimate is $E(P|X_1, \ldots, X_k)$. If it is desired to estimate the mean θ_0 of the distribution G with squared error loss a Bayes estimate is the mean of $E(P|X_1, \ldots, X_k)$, assuming it exists.

To compute $E(P|X_1, \ldots, X_k)$ notice that for $A \epsilon \mathcal{G}$

$$E[P(A)|X_1, \ldots, X_k] = E(E[P(A)|X_1, \ldots, X_k, \theta_1, \ldots, \theta_k]|X_1, \ldots, X_k)$$
$$= E(E[P(A)|\theta_1, \ldots, \theta_k]|X_1, \ldots, X_k)$$

$$= E\left[\frac{\alpha(A) + \delta_{\theta_1}(A) + \ldots + \delta_{\theta_k}(A)}{\alpha(\Theta) + k} \,\middle|\, X_1, \ldots, X_k\right]$$

$$= \frac{\alpha(A) + E[\delta_{\theta_1}(A) + \ldots + \delta_{\theta_k}(A) | X_1, \ldots, X_k]}{\alpha(\Theta) + k}$$

For $k = 1$ we get

$$E[P(A)|X_1] = \frac{\alpha(A) + H_{\theta_1|X_1}(A)}{\alpha(\Theta) + 1} .$$

To apply this technique to specific situations the problem becomes one of finding the distribution of the θ_i's conditional on the X_i's. In most cases this will be done by applying Bayes' theorem, so it will be necessary to investigate the joint distribution of $(\theta_1, \ldots, \theta_k)$. For the case $k = 2$, if $P \sim \mathcal{D}(\alpha)$ then it has been shown that $\theta_1 \sim \frac{\alpha(\cdot)}{\alpha(\Theta)}$ and $P|\theta_1 \sim \mathcal{D}(\alpha + \delta_{\theta_1})$, so $\theta_2|\theta_1 \sim \frac{\alpha(\cdot) + \delta_{\theta_1}(\cdot)}{\alpha(\Theta) + 1}$. In particular, if α has no points with positive mass then $P(\theta_1 = \theta_2) = \frac{1}{\alpha(\Theta) + 1} > 0$. The nature of the joint distribution of $(\theta_1, \ldots, \theta_k)$ makes the solution to the above statistical problems prohibitively complicated even for relatively small values of k; e.g. in the binomial case for $k = 4$ the estimated distribution function is a weighted sum of 52 distributions.

4. The binomial problem.

To treat the empirical Bayes problem when X_i has a binomial distribution with parameters n_i and θ_i for $i = 1, \ldots, k$, take $P \sim \mathcal{D}(MBe(a, b))$ where M is a positive constant and $Be(a, b)$ is the measure on $[0, 1]$ that has a beta distribution with parameters a and b. Consider the case $k = 2$, then

$$P|(X_1, X_2) \sim \int_{[0,1]^2} \mathcal{D}(MBe(a, b) + \delta_{\theta_1} + \delta_{\theta_2}) dH_{X_1, X_2}(\theta_1\ \theta_2),$$

$$\theta_1 \sim Be(a, b), \quad \theta_2|\theta_1 \sim \frac{MBe(a, b) + \delta_{\theta_1}}{M + 1}.$$

The joint distribution of (θ_1, θ_2) is a measure on $[0, 1]^2$ that is the weighted sum of the product of two beta measures with parameters a and b and a measure concentrated on the line $\theta_1 = \theta_2$ that has a $Be(a, b)$ distribution along the line. The respective weights are $P(\theta_1 \neq \theta_2) = \frac{M}{M + 1}$ and $P(\theta_1 = \theta_2) = \frac{1}{M + 1}$.

In finding the posterior distribution $H_{X_1, X_2}(\theta_1, \theta_2)$ it will be shown that Bayes' theorem can be applied separately to each part of the above measure and then the parts can be combined with their posterior weights $P(\theta_1 \neq \theta_2|X_1, X_2)$ and $P(\theta_1 = \theta_2|X_1, X_2)$. This will be shown in more generality than is needed here. For simplicity take X and θ as random variables. The proofs will hold with only minor modifications when X and θ are random vectors.

Assume $F_{X|\theta = y}$ exists and is a discrete probability distribution function almost surely dF_θ, where $F_{X|\theta = y}$ is by definition a function such that

$$P(X \leq x,\ \theta \epsilon S) = \int_S F_{X|\theta = y}(x) dF_\theta(y), \quad \text{for all } S \epsilon \mathcal{B}.$$

$F_{X|\theta = y}$ is absolutely continuous with respect to counting measure (denoted $m(\cdot)$) so that the Radon-Nikodym derivative $f_{X|\theta = y}$ exists. It is also assumed that $f_{X|\theta = y}(x)$ is a Borel-measurable function of y for all x.

We make two observations used in the proof of proposition one.

(1) $$F_X(x) = P(X \leq x) = \int_\Theta F_{X|\theta = y}(x) dF_\theta(y)$$

(2) $$F_{X|\theta = y}(x) = \int_{(-\infty, x]} f_{X|\theta = y}(t) dm(t).$$

Proposition 1: The measure corresponding to $F_X(x)$ is absolutely continuous with respect to counting measure and its Radon-Nikodym derivative $f_X(x)$ is

$$f_X(x) = \int_\Theta f_{X|\theta = y}(x) dF_\theta(y) \quad \text{a.s. } m$$

Proof:

$$F_X(x) = \int_\Theta F_{X|\theta = y}(x) dF_\theta(y) = \int_\Theta \int_{(-\infty, x]} f_{X|\theta = y}(t) dm(t) dF_\theta(y)$$

$$= \int_{(-\infty, x]} \int_\Theta f_{X|\theta = y}(t) dF_\theta(y) dm(t)$$

by Tonnelli's theorem.

Proposition 2: If the distribution function $F_{X|\theta = y}$ exists almost surely dF_θ then for any $S \in \mathcal{B}$ and $B \in \mathcal{B}$,

$$\int_B dF_{X|\theta = y}(x)$$

has meaning almost surely dF_θ, and

$$(3) \qquad P(\theta \in S,\ X \in B) = \int_S \int_B dF_{X|\theta = y}(x)\ dF_\theta(y).$$

Proof: We know that for any $a \in \mathbb{R}$

$$P(\theta \in S,\ X \in (-\infty,\ a]) = \int_S \int_{(-\infty, a]} dF_{X|\theta = y}(x) dF_\theta(y)$$

so we can apply the monotone class theorem (see [2], page 19) to get (3) for an arbitrary Borel measurable set B. We need to consider two cases.

Case 1: Let $B_n \nearrow B$ where for $n = 1, 2, \ldots, B_n$ is measurable and

$$P(\theta \in S,\ X \in B_n) = \int_S \int_{B_n} dF_{X|\theta = y}(x)\ dF_\theta(y).$$

We define disjoint sets $B_1', B_2', \ldots$ such that $B_1' = B_1$ and $B_n' = B_n - (\bigcup_{i=1}^{n-1} B_i)$ for $n = 2, 3, \ldots$. We show that (3) holds with $B = B_n'$. For B_1', the result is trivial. For n greater that 1.

$$P(\theta \epsilon S, X \epsilon B_n') = P(\theta \epsilon S, X \epsilon B_n) - P(\theta \epsilon S, X \epsilon B_{n-1})$$

$$= \int_S \int_{B_n} dF_{X|\theta = y}(x) dF_\theta(y) - \int_S \int_{B_{n-1}} dF_{X|\theta = y}(x) dF_\theta(y)$$

$$= \int_S \int_{B_n'} dF_{X|\theta = y}(x) dF_\theta(y).$$

Now, since $\bigcup_{i=1}^{m} B_i = \bigcup_{i=1}^{m} B_i'$ for $m = 1, 2, \ldots$,

$$P(\theta \epsilon S, X \epsilon B) = P(\theta \epsilon S, X \epsilon \bigcup_{n=1}^{\infty} B_n) = P(\theta \epsilon S, X \epsilon \bigcup_{n=1}^{\infty} B_n')$$

$$= \sum_{n=1}^{\infty} P(\theta \epsilon S, X \epsilon B_n') = \sum_{n=1}^{\infty} \int_S \int_{B_n'} dF_{X|\theta = y}(x) dF_\theta(y)$$

$$= \int_S \int_B dF_{X|\theta = y}(x) dF_\theta(y)$$

by monotone convergence.

Case 2: Let $B_n \nearrow B$ where for $n = 1, 2, \ldots$ B_n is measurable and

$$P(\theta \epsilon S,\ X \epsilon B_n) = \int_S \int_{B_n} dF_{X|\theta = y}(x)\ dF_\theta(y).$$

Now,

$$P(\theta \epsilon S,\ X \epsilon B) = \lim_{n\to\infty} P(\theta \epsilon S,\ X \epsilon B_n) = \lim_{n\to\infty} \int_S \int_{B_n} dF_{X|\theta = y}(x)\ dF_\theta(y).$$

By the dominated convergence theorem,

$$\lim_{n\to\infty} \int_S \int_{B_n} dF_{X|\theta = y}(x)\ dF_\theta(y) = \int_S \lim_{n\to\infty} \int_{B_n} dF_{X|\theta = y}(x)\ dF_\theta(y)$$

$$= \int_S \int_B dF_{X|\theta = y}(x)\ dF_\theta(y)$$

and the proof is complete.

The next two propositions are variations of Bayes' theorem.

Proposition 3: $F_{\theta|X = x}$ exists and

$$F_{\theta|X = x}(y) = \frac{\int_{(-\infty,y]} f_{X|\theta = t}(x)\ dF_\theta(t)}{\int_\Theta f_{X|\theta = t}(x)\ dF_\theta(t)}$$

almost surely dF_X.

Proof: Note that $\int_\Theta f_{X|\theta = t}(x)\ dF_\theta(t)$ is almost surely (dF_X) positive because it is the Radon-Nikodym derivative of F_X with respect to counting measure, so using proposition 1, take the exceptional set in this proposition as $\{x | f_X(x) = 0\}$.

Now, by proposition 2 and Tonnelli's theorem

$$P(X\epsilon B,\ \theta \le y) = \int_{(-\infty,y]} \int_B dF_{X|\theta = t}(x)\ dF_\theta(t)$$

$$= \int_{(-\infty,y]} \int_B f_{X|\theta = t}(x)\ dm(x)\ dF_\theta(t)$$

$$= \int_B \int_{(-\infty,y]} f_{X|\theta = t}(x)\ dF_\theta(t)\ dm(x).$$

Because the set $\{x|f_X(x) \neq 0\}$ has probability one,

$$\int_B \int_{(-\infty,y]} f_{X|\theta = t}(x)\ dF_\theta(t)\ dm(x)$$

$$= \int_B \frac{\int_{(-\infty,y]} f_{X|\theta = t}(x) dF_\theta(t)}{f_X(x)} f_X(x)\ dm(x)$$

$$= \int_B \frac{\int_{(-\infty,y]} f_{X|\theta = t}(x) dF_\theta(t)}{\int_\Theta f_{X|\theta = t}(x)\ dF_\theta(t)} dF_X(x).$$

So by the definition of $F_{\theta|X = x}$ the proposition holds.

For $S\epsilon \mathcal{B}$ such that $P(\theta\epsilon S|X = x) > 0$ almost surely dF_X define

$$F_{\theta|X = x, \theta\epsilon S}(y) = \frac{P((\theta \le y) \cap (\theta\epsilon S)|X = x)}{P(\theta\epsilon S|X = x)}$$

and

$$F_{\theta|\,\theta\epsilon S}(y) = \frac{P((\theta \le y) \cap (\theta\epsilon S))}{P(\theta\epsilon S)}.$$

Observe, for an S such that $0 < P(\theta\epsilon S|X=x) < 1$ a.s. dF_X

$$(4)\quad F_{\theta|X=x,\,\theta\epsilon S}(y)P(\theta\epsilon S|X=X) + F_{\theta|X=x,\,\theta\epsilon S^c}(y)P(\theta\epsilon S^c|X=x)$$

$$= F_{\theta|X=x}(y) \text{ a.s. } dF_X$$

where $S^c = \Theta - S$. Note also that

$$(5)\qquad F_{\theta|\,\theta\epsilon S}(y) = \int_{(-\infty,y]} dF_{\theta|\,\theta\epsilon S}(t) = \frac{P((\theta \le y) \cap (\theta\epsilon S))}{P(\theta\epsilon S)}$$

$$= \frac{1}{P(\theta\epsilon S)} \int_{(-\infty,y]} \chi_S(t)\, dF_\theta(t).$$

Proposition 4:

$$F_{\theta|X=x,\,\theta\epsilon S}(y) = \frac{\int_{(-\infty,y]} f_{X|\,\theta\,=\,t}(x)\, dF_{\theta|\,\theta\epsilon S}(t)}{\int_\Theta f_{X|\,\theta\,=\,t}(x)\, dF_{\theta|\,\theta\epsilon S}(t)}$$

almost surely dF_X.

Proof: By definition,

$$F_{\theta|X=x,\,\theta\epsilon S}(y) = \frac{P((\theta \le y) \cap (\theta\epsilon S)|X = x)}{P((\theta\epsilon S)|X = x)}$$

$$= \frac{\int_{(-\infty,y]\cap S} dF_{\theta|X=x}(t)}{\int_S dF_{\theta|X=x}(t)}$$

It is clear that a monotone class argument much like the one in proposition 2, gives for $A \in \mathcal{B}$,

$$\int_A dF_{\theta|X=x}(y) = \frac{\int_A f_{X|\theta=t}(x)dF_\theta(t)}{\int_\Theta f_{X|\theta=t}(x)dF_\theta(t)} \quad \text{a.s. } dF_X.$$

Now,

$$\frac{\int_{(-\infty,y]\cap S} dF_{\theta|X=x}(t)}{\int_S dF_{\theta|X=x}(t)} = \frac{\int_{(-\infty,y]\cap S} f_{X|\theta=t}(x)dF_\theta(t) \Big/ \int_\Theta f_{X|\theta=t}(x)dF_\theta(t)}{\int_S f_{X|\theta=t}(x)dF_\theta(t) \Big/ \int_\Theta f_{X|\theta=t}(x)dF_\theta(t)}$$

$$= \frac{\int_{(-\infty,y]\cap S} f_{X|\theta=t}(x)dF_\theta(t)}{\int_S f_{X|\theta=t}(x)dF_\theta(t)}$$

$$= \frac{\int_{(-\infty,y]} f_{X|\theta=t}(x)dF_{\theta|\theta\in S}(t)}{\int_\Theta f_{X|\theta=t}(x)dF_{\theta|\theta\in S}(t)}$$

The last equality follows from (5). The proposition is proven.

Proposition 4 and (4) give the major result. The last proposition shows how to compute the posterior weights in (4).

Proposition 5:

$$P(\theta\epsilon S|X=x) = \frac{P(\theta\epsilon S)\int_S f_{X|\theta=t}(x)dF_{\theta|\theta\epsilon S}(t)}{P(\theta\epsilon S)\int_\Theta f_{X|\theta=t}(x)dF_{\theta|\theta\epsilon S}(t)+P(\theta\epsilon S^c)\int_\Theta f_{X|\theta=t}(x)dF_{\theta|\theta\epsilon S^c}(t)}$$

for $S\epsilon\mathcal{B}$ such that $P(\theta\epsilon S|X=x)\epsilon(0, 1)$ almost surely dF_X.

Proof:

$$P(\theta\epsilon S|X=x) = \int_S dF_{\theta|X=x}(y) = \frac{\int_S f_{X|\theta=t}(x)dF_\theta(t)}{\int_\Theta f_{X|\theta=t}(x)dF_\theta(t)}$$

$$= \frac{P(\theta\epsilon S)\int_S f_{X|\theta=t}(x)dF_{\theta|\theta\epsilon S}(t)+P(\theta\epsilon S^c)\int_S f_{X|\theta=t}(x)dF_{\theta|\theta\epsilon S^c}(t)}{P(\theta\epsilon S)\int_\Theta f_{X|\theta=t}(x)dF_{\theta|\theta\epsilon S}(t)+P(\theta\epsilon S^c)\int_\Theta f_{X|\theta=t}(x)dF_{\theta|\theta\epsilon S^c}(t)}$$

$$= \frac{P(\theta\epsilon S)\int_S f_{X|\theta=t}(x)dF_{\theta|\theta\epsilon S}(t)}{P(\theta\epsilon S)\int_\Theta f_{X|\theta=t}(x)dF_{\theta|\theta\epsilon S}(t)+P(\theta\epsilon S^c)\int_\Theta f_{X|\theta=t}(x)dF_{\theta|\theta\epsilon S^c}(t)}.$$

Returning to the specific case at hand, Proposition 5 yields, in terms of odds,

$$\frac{P(\theta_1 = \theta_2 | X_1, X_2)}{P(\theta_1 \neq \theta_2 | X_1, X_2)}$$

$$= \frac{P(\theta_1=\theta_2)\dfrac{\Gamma(X_1+X_2+a)\Gamma(n_1+n_2-X_1-X_2+b)}{\Gamma(n_1+n_2+a+b)} \quad \dfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}}{P(\theta_1\neq\theta_2)\dfrac{\Gamma(X_1+a)\Gamma(n_1-X_1+b)\Gamma(X_2+a)\Gamma(n_2-X_2+b)\Gamma(a+b)\Gamma(a+b)}{\Gamma(n_1+a+b)\Gamma(n_2+a+b)\Gamma(a)\Gamma(b)\Gamma(a)\Gamma(b)}}$$

$$= \frac{a^{(X_1+X_2)}\, b^{(n_1+n_2-X_1-X_2)}}{(a+b)^{(n_1+n_2)}} \Bigg/ M\, \frac{a^{(X_1)}\, a^{(X_2)}\, b^{(n_1-X_1)}\, b^{(n_2-X_2)}}{(a+b)^{(n_1)}(a+b)^{(n_2)}}$$

where $a^{(n)} = a(a+1) \ldots (a+n-1)$.

Let $p_s \equiv P(\theta_1 = \theta_2 | X_1, X_2) \equiv 1 - p_d$, then

$$E(\theta_2 | X_1, X_2) = p_d \left(\frac{X_2+a}{n_2+a+b} \right) + p_s \left(\frac{X_1+X_2+a}{n_1+n_2+a+b} \right) ,$$

$$E(P | X_1, X_2) =$$

$$p_d \left[\frac{M}{M+2} \, Be(a, b) + \frac{1}{M+2} \, Be(X_1+a, n_1-X_1+b) \right.$$

$$\left. + \frac{1}{M+2} \, Be(X_2+a, n_2-X_2+b) \right]$$

$$+ p_s\left[\frac{M}{M+2}\,Be(a,b) + \frac{2}{M+2}\,Be(X_1+X_2+a,\ n_1+n_2-X_1-X_2+b)\right]$$

and the mean of $E(P|X_1, X_2)$ equals

$$p_d\left[\frac{M}{M+2}\left(\frac{a}{a+b}\right)+\frac{1}{M+2}\left(\frac{X_1+a}{n_1+a+b}\right)+\frac{1}{M+2}\left(\frac{X_2+a}{n_2+a+b}\right)\right]$$

$$+ p_s\left[\frac{M}{M+2}\left(\frac{a}{a+b}\right)+\frac{2}{M+2}\left(\frac{X_1+X_2+a}{n_1+n_2+a+b}\right)\right].$$

The extension for X_1, X_2, X_3 requires the following weights

$$P(\theta_1 \neq \theta_2 \neq \theta_3 \neq \theta_1|X_1, X_2, X_3) \equiv p_{1\cdot2\cdot3}$$

$$P(\theta_1 = \theta_2 \neq \theta_3|X_1, X_2, X_3) \equiv p_{12\cdot3}$$

$$P(\theta_1 = \theta_3 \neq \theta_2|X_1, X_2, X_3) \equiv p_{13\cdot2}$$

$$P(\theta_1 \neq \theta_2 = \theta_3|X_1, X_2, X_3) \equiv p_{23\cdot1}$$

$$P(\theta_1 = \theta_2 = \theta_3|X_1, X_2, X_3) \equiv p_{123}$$

The computation of these quantities is straightforward but tedious. The desired results are,

$$E(\theta_3|X_1, X_2, X_3) = p_{1\cdot2\cdot3}\left(\frac{X_3+a}{n_3+a+b}\right) + p_{12\cdot3}\left(\frac{X_3+a}{n_3+a+b}\right)$$

$$+ p_{13\cdot2}\left(\frac{X_1+X_3+a}{n_1+n_3+a+b}\right) + p_{23\cdot1}\left(\frac{X_2+X_3+a}{n_2+n_3+a+b}\right)$$

$$+ p_{123}\left(\frac{X_1 + X_2 + X_3 + a}{n_1 + n_2 + n_3 + a + b}\right)$$

and

$$E(P|X_1, X_2, X_3) = \frac{M}{M+3} Be(a, b)$$

$$+ p_{1\cdot2\cdot3}\left[\frac{1}{M+3} \sum_{i=1}^{3} Be(x_i + a, n_i - x_i + b)\right]$$

$$+ \sum p_{ij\cdot k}\left[\frac{2}{M+3} Be(x_i + x_j + a, n_i + n_j - x_i - x_j + b)\right.$$

$$\left. + \frac{1}{M+3} Be(X_k + a, n_k - X_k + b)\right]$$

$$+ p_{123}\left[\frac{3}{M+3} Be(X_1+X_2+X_3+a, n_1+n_2+n_3-X_1-X_2-X_3+b)\right]$$

where the sum is such that all the weights listed above are included once.

When $X_1, \ldots, X_4$ are observed there are 15 weighting factors and when $X_1, \ldots, X_5$ are observed there are 67.

Table 1 gives the estimate of θ_2 for several different observation pairs (X_1, n_1), (X_2, n_2) from various empirical Bayes estimators (see [6]). It also gives the posterior weights for the Ferguson - Antoniak estimate. Table 2 is similar to table 1 only it uses 3 observation pairs, estimates θ_3, and does not give posterior weights.

TABLE 1
EMPIRICAL BAYES ESTIMATES

ESTIMATE OF θ_2 (X_1,n_1) (X_2,n_2)	(4,5) (9,10)	(1,5) (8,10)	(3,5) (7,10)	(1,5) (9,10)	(17, 19) (28, 29)	(1, 19) (28, 29)	(10, 19) (21, 29)
MAXIMUM LIKELIHOOD	.900	.800	.700	.900	.966	.966	.724
POOLED	.867	.600	.667	.667	.938	.604	.646
MINIMAX	.804	.728	.652	.804	.893	.893	.689
COPAS'S 1st	.888	.728	.688	.816	.960	.894	.709
COPAS'S 2nd	.850	.730	.650	.855	.930	.962	.625
Griffin & Krutchkoff	.900	.741	.700	.860	.966	.962	.643
Ferguson - Antoniak $a = b = M = 1$	.827	.725	.654	.821	.923	.935	.673
$P(\theta_1=\theta_2 \mid (X_1,n_1),(X_2,n_2)$	.663	.156	.622	.0643	.778	5.85×10^{-10}	.534

TABLE 2

EMPIRICAL BAYES ESTIMATES

(X_1,n_1) (X_2,n_2) (X_3,n_3) ESTIMATE OF θ_3	(2, 5) (3, 5) (2, 5)	(1, 5) (3, 10 (2, 5)	(3, 5) (7, 10) (4, 10)	(1, 5) (9, 10) (8, 10)	(9, 10) (10, 12) (6, 12)	(9, 10) (1, 12) (10, 12)	(9, 10) (10, 12) (1, 12)
MAXIMUM LIKELIHOOD	.400	.400	.400	.800	.500	.833	.083
POOLED	.467	.300	.560	.720	.735	.588	.588
MINIMAX	.431	.431	.424	.728	.500	.759	.177
COPAS'S 1st	.421	.369	.440	.760	.555	.782	.200
COPAS'S 2nd	.467	.300	.567	.771	.620	.818	.119
Griffin & Krutchkoff	.400	.400	.400	.774	.616	.818	.119
Ferguson - Antoniak a = b = M = 1	.455	.351	.494	.780	.572	.820	.143
Ferguson - Antoniak a = M = .5 b = 2	.414	.296	.499	.746	.633	.793	.104

Acknowledgements

The author would like to express his appreciation to Donald Berry who introduced him to the field and whose suggestions were very beneficial.

References

[1] Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Ann. Statist., 2, 1152-1174.

[2] Ash, R. B. (1972). Real Analysis and Probability. New York: Academic Press.

[3] Breiman, L. (1968). Probability. Reading, Mass.: Addison-Wesley.

[4] Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. Ann. Statist., 1, 209-230.

[5] Maritz, J. S. (1970). Empirical Bayes Methods. London: Methuen.

[6] Martz, H. F. and Lian, M. G. (1974). Empirical Bayes estimation of the binomial parameter. Biometrika, 61, 517-523.